Exact and Approximate Bayesian Inference for Low Count Time Series Models with Intractable Likelihoods

C. C. Drovandi and A. N. Pettitt
Mathematical Sciences
Queensland University of Technology, Brisbane, Australia 4000
email: c.drovandi@qut.edu.au

May 15, 2013

Abstract

In this paper we present a new simulation methodology in order to obtain exact or approximate Bayesian inference for models for low-valued count time series data that have computationally demanding likelihood functions. The algorithm fits within the framework of particle Markov chain Monte Carlo (PMCMC) methods. The particle filter requires only model simulations and, in this regard, our approach has connections with approximate Bayesian computation (ABC). However, an advantage of using the PMCMC approach in this setting is that simulated data can be matched with data observed one-at-a-time, rather than attempting to match on the full dataset simultaneously or on a low-dimensional non-sufficient summary statistic, which is common practice in ABC. For low-valued count time series data we find that it is often computationally feasible to match simulated data with observed data exactly. Our particle filter maintains N particles by repeating the simulation until $N+1$ exact matches are obtained. Our algorithm creates an unbiased estimate of the likelihood, resulting in exact posterior inferences when included in an MCMC algorithm. In cases where exact matching is computationally prohibitive, a tolerance is introduced as per ABC. A novel aspect of our approach is that we introduce auxiliary variables into our particle filter so that partially observed and/or non-Markovian models can be accommodated. We demonstrate that Bayesian model choice problems can be easily handled in this framework.

Keywords: Approximate Bayesian computation, INARMA model, Markov process, particle filter, particle marginal Metropolis-Hastings, particle Markov chain Monte Carlo, renewal process.

1 Introduction

In this paper a new simulation methodology is presented to perform exact and approximate Bayesian inference (the latter is referred to as approximate Bayesian computation (ABC)) on model parameters for data of low count time series. The approach relies on the simulation from the likelihood in order to avoid likelihood evaluations, which can be cumbersome or even

intractable for some count time series models in the literature (e.g. Markov process, integer autoregressive moving average (INARMA) and renewal process models). Models are not all Markov, and can have long memory such as in the case of a renewal process model. In some cases the likelihood is itself approximated.

In the case of exact Bayesian inference, the simulated likelihood for a parameter value is obtained efficiently by matching simulated data on the observed data sequentially via a particle filter, which is typically straightforward to do for low count time series data. The simulated likelihood obtained from the particle filter is incorporated within a Metropolis-Hastings algorithm, so obtaining the particle Markov chain Monte Carlo (PMCMC) approach of Andrieu et al. (2010). More specifically, we use the particle marginal Metropolis-Hastings (PMMH) approach of Andrieu et al. (2010). We note that an MH algorithm which uses an unbiased estimator of the likelihood produces an exact algorithm in the sense that the appropriately marginalised target distribution is still the true posterior distribution (Andrieu and Roberts, 2009).

For data and/or models where exact matching is computationally prohibitive, one can assume that the observed data is measured with error so obtaining a state-space model that approximates the true model (see Wilkinson (2008)). The non-exact matching case is also referred to as ABC. We again use the PMMH algorithm, however the target distribution here is an ABC target, referred to here as approximate Bayesian inference. Regardless of the matching strategy, we are always using the full data so avoiding the need to choose a summary statistic (see Blum et al. (2012)).

In most applications of PMCMC, the data has a Markov structure. However, we amend the particle filter to carry through auxiliary information in order to handle non-Markovian models. In addition, this formulation can be used to accommodate partially observed data (that is, one or more of the variables in the full model are unobserved), which is common in epidemic modelling (see, for example, Drovandi and Pettitt (2008)) and systems biology (see, for example, Toni et al. (2009)).

We apply the methodology to a number of examples. The first example is a multivariate Markov process for epidemic data (see O'Neill and Roberts (1999)). This application is mainly for illustrative purposes since for low-valued discrete counts the exact likelihood is computationally feasible using the matrix exponential (Moler and van Loan, 2003; Sidje, 1998). However, with these types of models the likelihood computation grows exponentially as more random variables are added to the model (see, for example, Drovandi and Pettitt (2008)). O'Neill and Roberts (1999) complete the likelihood with missing information and select priors to permit closed form full conditional sampling.

The second type of model is the integer autoregressive moving average (INARMA) model. The INAR(p) model with order $p = 1$ has a trivial likelihood, but becomes computationally intensive for higher orders. White et al. (2013) present a Bayesian methodology to handle Markovian models such as this one and discretely observed Markov processes. However, the methodology is restricted to models that are Markovian, which rules out application to INARMA models with a moving average component. Furthermore, for multivariate models all variables must be observed. Neal and Subba Rao (2007) develop a component-wise Bayesian MCMC algorithm for INARMA models by completing the likelihood with auxiliary variables. This is extended by Enciso-Mora et al. (2009) to a reversible jump (RJ) MCMC algorithm for

model selection between competing INARMA models. Our algorithm is on the marginal space of the parameter of interest and can therefore avoid potentially poor mixing of the MCMC sampler on the joint space of parameter and latent variables. This was one of the original motivations for the PMCMC algorithms and the pseudo marginal approach of Andrieu and Roberts (2009). Furthermore, we can develop a more efficient RJMCMC algorithm on the marginal space.

The third and final application of our methodology is to renewal process models, see, for example, Cui and Lund (2009). The technique of Cui and Lund (2009) superpositions several independent copies of a renewal process to create correlated count time series with a specified marginal distribution. The renewal process is defined by a so-called lifetime distribution, and very flexible renewal processes can be obtained with only a small number of parameters, including generating sequences with long memory. Cui and Lund (2009) propose an estimation approach which is optimal for Gaussian short memory time series and is inefficient for non-Gaussian time series or long memory series. Here we use our algorithm to perform exact and approximate Bayesian inference on these models.

This paper is organised as follows. In Section 2 the algorithm is presented. Section 3 considers the examples specified above. Section 4 contains the discussion together with the limitations of the algorithm.

2 Methods

2.1 Notation

We denote the observed data (possibly vector) at time $t \in \{1, \ldots, T\}$ as $\boldsymbol{y_t}$ where T is the number of observations. The associated auxiliary information in the particle filter (when required) is given by $\boldsymbol{x_t}$. The parameter of the model is $\boldsymbol{\theta}$ with prior density $p(\boldsymbol{\theta})$. The likelihood, $p(\boldsymbol{y}|\boldsymbol{\theta})$, where $\boldsymbol{y} = (\boldsymbol{y_1}, \ldots, \boldsymbol{y_T})$ is combined with the prior, $p(\boldsymbol{\theta})$, to produce the posterior, $p(\boldsymbol{\theta}|\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})/Z$ where $Z = p(\boldsymbol{y})$ is the normalising constant.

2.2 The Algorithm

The algorithm is performed on the marginal space of $\boldsymbol{\theta}$, as opposed to the joint space of $\boldsymbol{\theta}$ and any auxiliary variables, in the same vein as PMCMC. Andrieu and Roberts (2009) note that the Metropolis-Hastings algorithm still targets the posterior, $p(\boldsymbol{\theta}|\boldsymbol{y})$, as long as the likelihood is estimated unbiasedly. The simplest example of this is the simulated likelihood, where the probability of generating the data for a particular $\boldsymbol{\theta}$ is estimated via simulation. Therefore the use of the simulated likelihood in an MCMC algorithm for posterior simulation can still produce exact inferences (in the sense that the algorithm still targets the marginal posterior distribution of $\boldsymbol{\theta}$, see Flury and Shephard (2011) for an application of this) if the simulated likelihood is estimated unbiasedly. Unfortunately the same exactness is not necessarily achieved for the simulated maximum likelihood approach (see Cappellari and Jenkins (2003) for an application of simulated maximum likelihood).

Our implementation of the particle filter maintains exactly N particles throughout the particle

filter via negative binomial resampling until $N+1$ 'matches' are obtained (the rationale for requiring $N+1$ matches is provided below). This type of approach is suggested by Le Gland and Oudjane (2006) in order to avoid degeneracy in the particle filter. The method is referred to as the alive particle filter in Jasra et al. (2013). The natural implementation of this type of filter (resampling until N matches, as done in Le Gland and Oudjane (2006)) results in a biased estimate of the likelihood. However, Jasra et al. (2013) propose a correction to produce an unbiased estimate, which is attractive when using the resulting estimate within an MCMC algorithm (Andrieu and Roberts, 2009). We apply this correction in our paper. The correction involves producing $N+1$ matches and not including the $N+1$th match in the particle set. The likelihood component for a particular observation is estimated by dividing N by the total number of simulations required to produce $N+1$ matches minus one.

If the model is Markov and exact matching of the data is not too computationally intensive, the simulated likelihood can be computed. Note that the Markovian structure in the likelihood means that exact matching is required only for one observation at a time conditional on the previous observations (depending on the Markovian order) being fixed. From a Bayesian point of view, the approach of White et al. (2013) is applicable to such Markovian models. However, if one or more of the variables (or populations from an epidemic perspective) is unobserved, the method of White et al. (2013) is not applicable in its current form. Furthermore the method can only handle models that are Markovian. We require some alternative to the simulated likelihood if:

a exact matching is too computationally intensive,

b one or more of the populations in the model is unobserved, and/or

c the model is non-Markovian.

In the case of **a**, exact Bayesian inference can be replaced with ABC by being less stringent on the matching. For low-valued discrete counts, a reasonable approximate target distribution is given by

$$\begin{aligned} p_\epsilon(\boldsymbol{\theta}, \boldsymbol{s}|\boldsymbol{y}) &\propto p(\boldsymbol{\theta})p(\boldsymbol{s}|\boldsymbol{\theta}) \prod_{t=1}^{T} 1(||\boldsymbol{s_t} - \boldsymbol{y_t}|| \leq \epsilon_t) \\ &= p(\boldsymbol{\theta})p(\boldsymbol{s_1}|\boldsymbol{\theta})1(||\boldsymbol{s_1} - \boldsymbol{y_1}|| \leq \epsilon_1) \prod_{t=2}^{T} p(\boldsymbol{s_t}|\boldsymbol{s_1}, \ldots, \boldsymbol{s_{t-1}}, \boldsymbol{\theta})1(||\boldsymbol{s_t} - \boldsymbol{y_t}|| \leq \epsilon_t), \end{aligned}$$

where $\boldsymbol{s} = (\boldsymbol{s_1}, \ldots, \boldsymbol{s_T})$ is the auxiliary simulated data on the same support as the observed data and ϵ_t are the tolerances. For time series of low counts, a reasonable discrepancy function is the L_1-norm (the sum of the absolute differences between every component of the observed and simulated data, since $\boldsymbol{y_t}$ is possibly vector-valued). We allow different tolerances for different observations since there may be heteroscedasticity or outliers in the data. Assuming that the model is Markovian and all variables in the model are observed, this is effectively a simulated ABC likelihood, which gives an approximation to the true ABC likelihood which is given by

$$p_\epsilon(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{\boldsymbol{s_1}} \ldots \sum_{\boldsymbol{s_T}} p(\boldsymbol{s_1}|\boldsymbol{\theta})1(||\boldsymbol{s_1} - \boldsymbol{y_1}|| \leq \epsilon_1) \prod_{t=2}^{T} p(\boldsymbol{s_t}|\boldsymbol{s_1}, \ldots, \boldsymbol{s_{t-1}}, \boldsymbol{\theta})1(||\boldsymbol{s_t} - \boldsymbol{y_t}|| \leq \epsilon_t).$$

Under this situation, the algorithm targets the ABC posterior; our simulated ABC likelihood is an unbiased estimate of the true ABC likelihood. An alternative perspective is that exact Bayesian inference is obtained under the assumption of observation error (Wilkinson, 2008) with ϵ_t given a discrete distribution.

Under the conditions of **a**, **b** or **c** above, a particle filter can be used to obtain an unbiased estimate of the true or ABC likelihood for a particular value of $\boldsymbol{\theta}$. For models with computationally difficult likelihoods, an attractive particle filter is the bootstrap particle filter (Gordon et al., 1993). In this particle filter, only sequential simulation of the data is required and no transition probabilities need to be evaluated. We make use of auxiliary information, $\boldsymbol{x_t}$, in our particle filter, to account for the possibility of **b** and/or **c** mentioned above. The make-up of the auxiliary variable is problem specific; more details are provided in the examples section. At each iteration of the particle filter, we keep track of an approximation to the posterior of the auxiliary data, $p(\boldsymbol{x_t}|\boldsymbol{s_1}, \ldots, \boldsymbol{s_t}, \boldsymbol{\theta}) \prod_{j=1}^{t} 1(||\boldsymbol{s_j} - \boldsymbol{y_j}|| \leq \epsilon_j)$. The auxiliary data may carry the values of unobserved random variables or information that may be required to simulate sequentially non-Markovian models. Both of these aspects will become clear in the examples in Section 3. It is worth noting that even though our particle filter makes use of auxiliary variables, it is only used to obtain the simulated likelihood, which is then used in the MCMC sampler on the marginal space of $\boldsymbol{\theta}$.

The MCMC algorithm on the marginal space is shown in Algorithm 1, while the workhorse part of the method, which we refer to as the alive bootstrap particle filter with auxiliary variables, is shown in Algorithm 2. The use of MCMC is critical for this approach: for a good proposal distribution, it ensures that we are sampling in regions where the posterior is non-negligible. Sequential Monte Carlo methods that begin from a vague distribution relative to the posterior may not be successful in this setting. Poor parameter proposals may never be able to generate simulated data close to the observed data, even one-at-a-time as we do here.

Algorithm 1 PMCMC algorithm (PMMH algorithm of Andrieu et al. (2010)) to simulate from the ABC target. When $\epsilon = 0$ exact inferences are obtained.

Input: $\boldsymbol{\theta^0}$ and iters
Output: MCMC output $\boldsymbol{\theta^1}, \ldots, \boldsymbol{\theta^{iters}}$

Compute $\phi^0 = \hat{p}_\epsilon(\boldsymbol{y}|\boldsymbol{\theta^0})$ (using Algorithm 2)
for $i = 1$ **to** iters **do**
  Propose $\boldsymbol{\theta^*} \sim q(\cdot|\boldsymbol{\theta^{i-1}})$
  Compute $\phi^* = \hat{p}_\epsilon(\boldsymbol{y}|\boldsymbol{\theta^*})$ (using Algorithm 2)
  Compute $\alpha = \min\left(1, \frac{\phi^* p(\boldsymbol{\theta^*}) q(\boldsymbol{\theta^{i-1}}|\boldsymbol{\theta^*})}{\phi^{i-1} p(\boldsymbol{\theta^{i-1}}) q(\boldsymbol{\theta^*}|\boldsymbol{\theta^{i-1}})}\right)$
  if $U(0,1) < \alpha$ **then**
    Set $\phi^i = \phi^*$ and $\boldsymbol{\theta^i} = \boldsymbol{\theta^*}$
  else
    Set $\phi^i = \phi^{i-1}$ and $\boldsymbol{\theta^i} = \boldsymbol{\theta^{i-1}}$
  end if
end for

There are a few remarks to be made about Algorithm 2. Firstly if exact matching is performed, the variables $\boldsymbol{s_t}$, clearly do not need to be maintained in the algorithm since it will be equal

Algorithm 2 The alive bootstrap particle filter with auxiliary variables. When $\epsilon = 0$ the algorithm produces an estimate of the true likelihood. When the process is Markovian and exact matching can be performed on all observed variables, the algorithm is technically not a particle filter and line 8 is not required.

Input: A particular value of the parameter, $\boldsymbol{\theta}$, and the number of particles, N
Output: $\log \hat{p}_\epsilon(\boldsymbol{y}|\boldsymbol{\theta})$ (i.e. the log of the estimated ABC likelihood)

1: Set $\log \hat{p}_\epsilon(\boldsymbol{y}|\boldsymbol{\theta}) = 0$
2: Obtain initial simulated data, $\boldsymbol{s_0^i}$, and auxilary variable information, $\boldsymbol{x_0^i}$, for $i = 1, \ldots, N$
3: **for** $t = 1$ **to** T **do**
4: Set sims=0
5: **for** $k = 1$ **to** $N + 1$ **do**
6: matched = 'no'
7: **while** matched == 'no' **do**
8: Resample an index r from the set $\{1, \ldots, N\}$ with equal weights via multinomial resampling
9: Generate $\boldsymbol{s_t^*}$ and $\boldsymbol{x_t^*}$ from the model based on $\boldsymbol{s_{t-1}^r}$ and $\boldsymbol{x_{t-1}^r}$
10: Set sims = sims+1
11: **if** $||\boldsymbol{s_t} - \boldsymbol{y_t}|| \leq \epsilon_t$ **then**
12: Set $\boldsymbol{s_t^k} = \boldsymbol{s_t^*}$, $\boldsymbol{x_t^k} = \boldsymbol{x_t^*}$ and matched = 'yes'
13: **end if**
14: **end while**
15: **end for**
16: Set $\log \hat{p}_\epsilon(\boldsymbol{y}|\boldsymbol{\theta}) = \log \hat{p}_\epsilon(\boldsymbol{y}|\boldsymbol{\theta}) + \log(N/(\text{sims} - 1))$
17: **end for**

to the data for all particles. If the model is Markovian and all variables are observed, the auxiliary information, $\boldsymbol{x_t}$, is also not required. If both of these are true then technically the algorithm is not a particle filter, as no information needs to be maintained sequentially. The resample step is obviously no longer required.

For a very poor parameter proposal, the model may have very little chance of generating simulated data close to observed data and the particle filter could take an excessively long time or could get stuck. We suggest implementing an intervention in the algorithm which checks the value of 'sims' in Algorithm 2 and if it becomes excessively large then reject that value of $\boldsymbol{\theta}$ in Algorithm 1. This will create a very small amount of bias.

The algorithm is based on model simulation and hence is somewhat plug-and-play. However, there are some problem specific implementation considerations such as specifying the required auxiliary variables. This becomes more clear in the examples section below. The computational cost of the algorithm is mostly consumed within the particle filter. Therefore we suggest using a low level implementation (e.g. C, Fortran) of this part of the algorithm.

3 Examples

3.1 Epidemic Models

Here we apply the methodology using the general epidemic stochastic model of O'Neill and Roberts (1999) to the inter-removal times data based on a smallpox epidemic in Abakaliki, Nigeria (see Becker (1989, pg. 111)). The village consists of 120 people, one of whom becomes infected with smallpox and introduces it into the community. The observation of the process begins upon the first removal, which is set at time $t = 0$. The infection times and the number of susceptibles are unobserved throughout the process, apart from the end of the epidemic at which point there are no infectives. Before this there must always be at least one infective so that the epidemic can continue. The motivation for this example is to demonstrate the methodology on partially observed processes. The exact likelihood for this partially observed Markov process could be computed using the approach of Drovandi and Pettitt (2008) based on the matrix exponential.

We denote the number of susceptibles, infectives and removals at time t as S_t, I_t and R_t, respectively, with $R_0 = 1$. In an infinitesimal time Δ_t a transmission can occur with probability $\beta S_t I_t \Delta_t$, which increments the number of infectives and reduces the number of susceptibles by one. A removal can occur in that time with probability $\gamma I_t \Delta_t$. Removed individuals play no further part in the process. The removals data is shown in Figure 1. The epidemic ceases at day 76. O'Neill and Roberts (1999) and Bailey (1975) mention that the model is not particularly realistic for a smallpox epidemic, however we apply the model for illustrative purposes.

O'Neill and Roberts (1999) apply a Bayesian MCMC algorithm that includes the missing infection times as auxiliary variables. Our sampler is on the marginal space and thus should mix more readily. O'Neill and Roberts (1999) conveniently choose gamma priors to permit closed-form full conditional sampling. Here we run the algorithm with two priors: (1) the prior of O'Neill and Roberts (1999), that is the first infection time $t < 0$ has an uninfor-

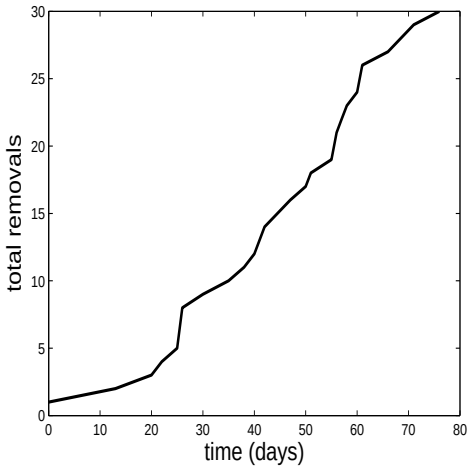


Figure 1: Smallpox epidemic data on removals in Abakaliki, Nigeria. Data source: Becker (1989, pg. 111)

mative improper prior and the priors for δ and β are improper and each is proportional to the reciprocal of the parameter; (2) is the same as (1) except the priors for δ and β are Uniform$(0, \infty)$.

The numbers of susceptibles and infectives at each observed removal time are auxiliary variables in the particle filter ($\boldsymbol{x_t} = (S_t, I_t)$), with $N = 200$ particles. We use the following data matching scheme for our approach. Initially the epidemic is simulated for each particle until the first removal, conditional on there being at least one infective at $t = 0$. Subsequently the number of removals are matched exactly at each observed removal time. Again we condition on $I_t > 0$ except for $I_{76} = 0$. Hence $\boldsymbol{y_t} = (R_t, I_t > 0)$ for $t < 76$ and $\boldsymbol{y_{76}} = (R_{76}, I_{76} = 0)$.

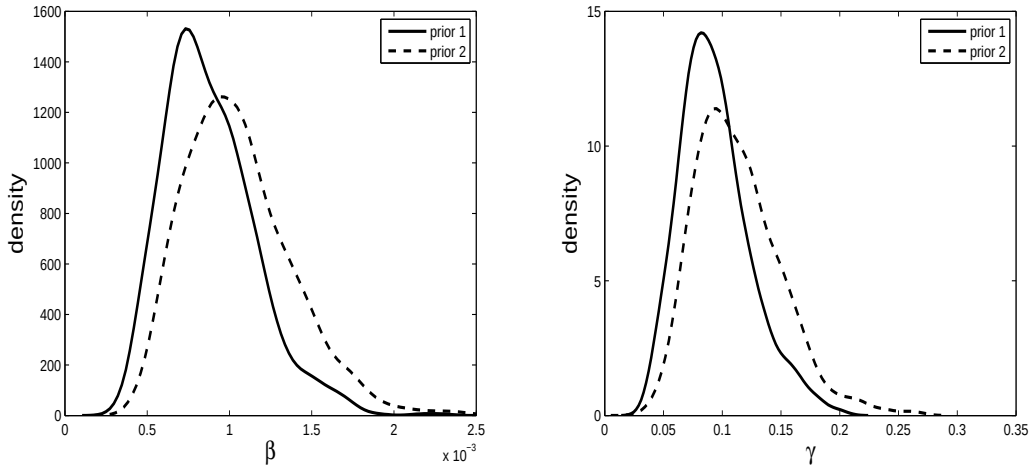
The posterior distributions for both priors are shown in Figure 2 for β (Figure 2(a)) and γ (Figure 2(b)). The posteriors obtained for prior 1 appear to be in agreement with Figure 4 of O'Neill and Roberts (1999). The posteriors are somewhat sensitive to the choice of the priors.

3.2 Integer Autoregressive Moving Average Models

The integer autoregressive moving average (INARMA) model is the discrete version of the popular ARMA model for stationary Gaussian time series. The INARMA(p, q) model is given by

$$Y_t = \sum_{i=1}^{p} \alpha_i \circ Y_{t-i} + Z_t + \sum_{j=1}^{q} \beta_j \circ Z_{t-j},$$

where $\circ$ is the binomial thinning operator (that is, if $W = \alpha \circ Y$, then $W \sim \text{Binomial}(Y, \alpha)$) and Z_t for $t \in \mathcal{N}$ is a sequence of independent and identically distributed discrete random variables. A popular choice is $Z_t \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$, which is adopted here. The likelihood is cumbersome for all but the INAR(1) model, which involves the convolution of a binomial and Poisson random variable. From a Bayesian perspective, such models have been studied via



(a) posterior for β (b) posterior for γ

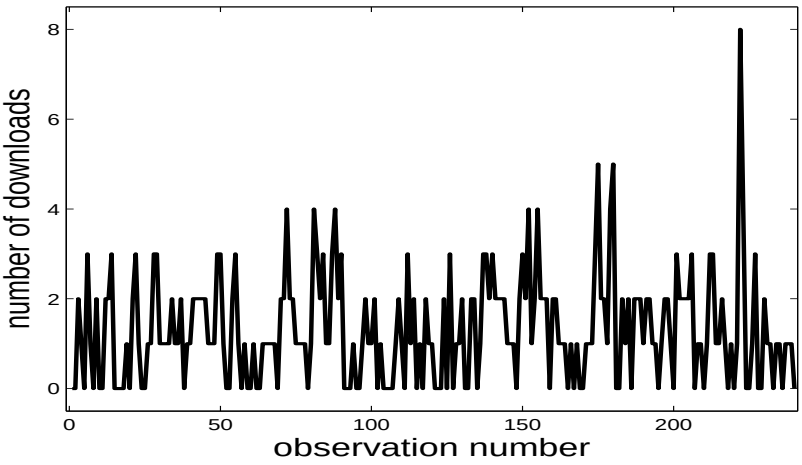
Figure 2: Posterior densities for when the general epidemic model is applied to the smallpox data in Figure 1. Solid denotes the posterior based on prior 1 and dashed the posterior based on prior 2.

introducing auxiliary variables to form the complete likelihood and using MCMC for joint posterior simulation and RJMCMC for selecting the model order. Our method samples over the marginal space of $\boldsymbol{\theta} = (p, \alpha_1, \ldots, \alpha_p, q, \beta_1, \ldots, \beta_q)$ only. The approach of White et al. (2013) can handle INAR(p) models but cannot accommodate a moving average component due to the lack of Markov structure in the resulting model. Our particle filter uses auxiliary variables to allow the addition of the moving average component.

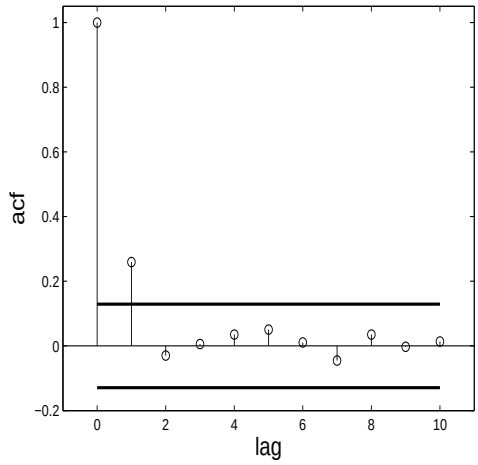
We analyse the number of web address downloads at a computer at the University of Würzburg (see Martin et al. (2011)). From the data, autcorrelation function and partial autocorrelation function, it is evident that the INAR(1), INMA(1) and INARMA(1,1) are all plausible models. However, none of the models are able to replicate the observed value of 8 in this dataset, suggesting that this point is an outlier with respect to these models (see Eduarda Silva and Pereira (2012) for confirmation of this). This observation was changed to a 5 for our purposes, effectively setting $\epsilon = 3$ for this observation.

For all models the parameters on (0,1) had a Uniform$(0, 1)$ prior whilst the λ parameter had an improper uniform prior on $\mathbb{R}^+$, $\lambda \sim$ Uniform$(0, \infty)$. The algorithm for the models INAR(1) and INMA(1) was run for 21000 iterations whilst the INARMA(1,1) for 31000 iterations, discarding the first 1000 as burn-in in all cases. Multivariate normal random walks were used with appropriate tuning parameters.

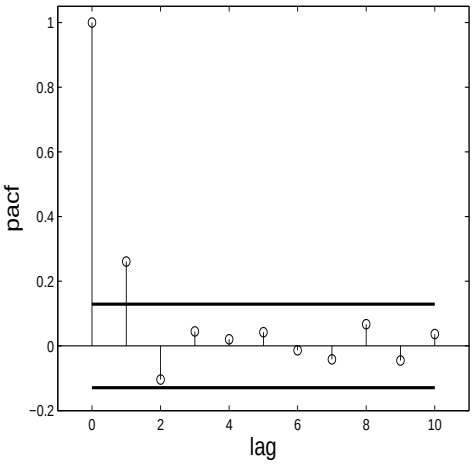
Normal distributions were fitted to the posterior samples of the INAR(1) and INMA(1) model parameters. Gamma distributions were fitted for α and β while a normal distribution was fitted for λ for the INARMA(1,1) model. These proposals were used in the reversible jump algorithm. At each iteration a proposal was made to one of the other models with equal probability. The use of an improper prior for λ is valid here as we assume the same arbitrary constant for each model and thus have cancelling in the marginal likelihood ratio (Pericchi,



(a) data



(b) acf



(c) pacf

Figure 3: (a) The number of web address downloads every 2 minutes. (b) Autocorrelation function. (c) Partial autocorrelation function. Data source: Martin et al. (2011).

2005). The RJMCMC was run for 50000 iterations with an appropriate starting value (so no burn-in was required). The posterior probability of the INAR(1) and INMA(1) models was roughly 0.41 while the INARMA(1,1) had probability 0.18. Posterior samples for the parameters were also drawn from this run, using a thinning factor of 20. These posterior distributions can be seen in Figure 4. Another example on INARMA models is provided in Appendix A.

We also investigated a random co-efficient integer autoregressive model of order 1 (RCINAR(1)). In this example the model is given by

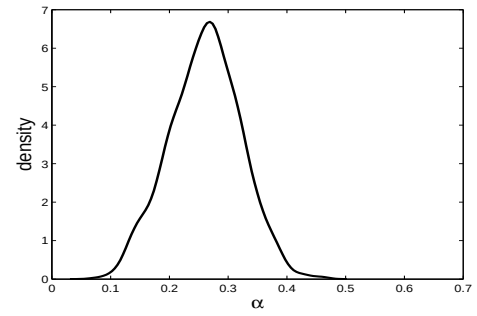
$$Y_t = \alpha_t \circ Y_{t-1} + Z_t,$$

where $\alpha_t \sim \text{Uniform}(0, 2\phi)$ and $Z_t \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$. Zhang et al. (2011) propose an empirical likelihood approach. Our approach uses the true likelihood. 50 observations were simulated with $\phi = 0.2$, $\lambda = 1.5$ and $Y_0 = 3$ (we assume that Y_0 is fixed for simplicity). Here $\phi \sim \text{Uniform}(0, 0.5)$ and $p(\lambda) \propto 1(\lambda > 0)$. In our approach, for each proposed simulation for each particle, an α_t is drawn independently. We used $N = 100$ particles and a suitable joint proposal for (ϕ, λ) in the marginal MCMC algorithm. To validate our algorithm, we compared it with a Bayesian approach where the parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{50})$ is introduced to form a complete likelihood. This algorithm uses the same proposal for ϕ and λ then draws the vector $\boldsymbol{\alpha}$ independently based on the proposed ϕ. Our acceptance probability involves the ratio of simulated likelihoods whilst the other Bayesian algorithm computes the ratio of the complete likelihoods (convolution of a binomial and Poisson random variable). Both posterior distributions produced were very similar. With the particle filter being implemented in C (and with a Matlab implementation elsewhere), both algorithms required roughly the same computer time per iteration. The marginal algorithm had an acceptance rate of 40%, whilst being only 12% for the joint sampler. After thinning both outputs to give the same number of samples, the marginal sampler produced a lower autocorrelation.

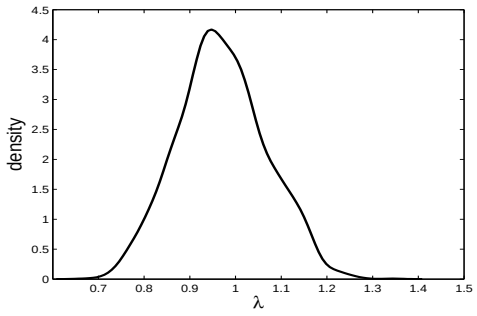
3.3 Renewal Process Models

Finally the methodology is presented on the most challenging model thus far, which involves running a number of independent renewal processes in parallel. These models are developed by Cui and Lund (2009). However, we provide the details here for completeness. One must specify the mass function of a random variable, $P(L = n) = f_n$ with $n \geq 1$, often referred to as the lifetime distribution. Consider a sequence of random variables $L_0, L_1, L_2, \ldots$, with L_i for $i \geq 1$ independent and identically distributed copies of L and L_0 is allowed its own distribution. In this paper we consider $L_0 = 0$ giving a non-delayed renewal process. A renewal is set to take place at integer time t if $L_1 + L_2 + \cdots + L_k = t$ for some $k \geq 1$. Consider a sequence of binary random variables, $A_1, A_2, \ldots$. If a renewal takes place at time t then $A_t = 1$ otherwise $A_t = 0$.

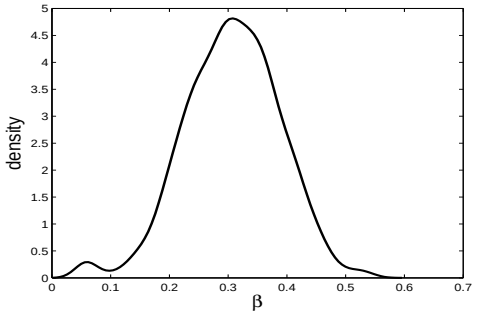
Any discrete distribution on the non-negative integers can be realised by considering the sum of a possibly random number of binary random variables. Therefore these types of models can create correlated discrete time series with specified marginals by superpositioning several independent copies of the renewal process. For an example, the sum of m (fixed) independent renewal processes creates a correlated sequence of discrete random variables with binomial



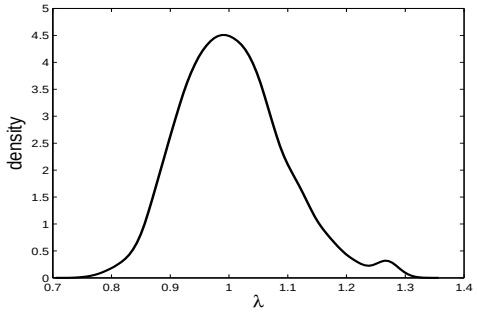
(a) α, INAR(1)



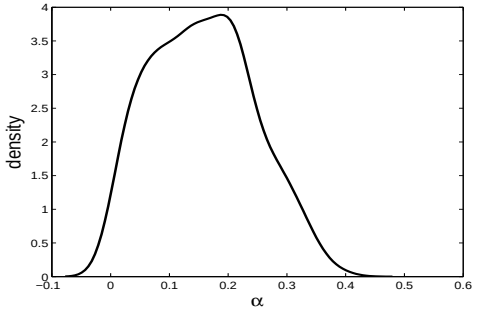
(b) λ, INAR(1)



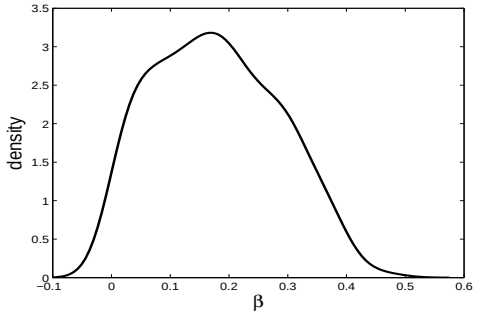
(c) β, INMA(1)



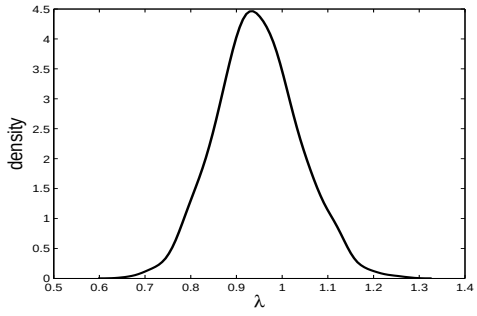
(d) λ, INMA(1)



(e) α, INARMA(1,1)



(f) β, INARMA(1,1)



(g) λ, INARMA(1,1)

Figure 4: Posterior densities for when the INAR(1), INMA(1) and INARMA(1,1) models are fitted to the downloads data in Figure 3. (a) and (b) Posteriors for INAR(1) model. (c) and (d) Posteriors for INAR(2) model. (e), (f) and (g) Posteriors for INARMA(1,1).

marginals. That is, $Y_t = A_{1,t} + A_{2,t} + \cdots + A_{m,t}$, where $A_{j,t} \in \{0, 1\}$ is the value of the jth renewal process at time t.

The flexibility of these distributions comes with the lifetime distribution choice for the random variable L. Lifetime distributions can be specified to create a process that is non-Markovian or even long memory (see below for an example). In general the likelihood function is intractable. Cui and Lund (2009) propose an estimation methodology that minimises the sum of the squares of the differences between the data and the best linear predictor based on the past history of the data. This approach may be inefficient and even inappropriate for long memory time series.

Implementing our method for this model requires the use of auxiliary variables in the particle filter. Let $x_{i,t} = L_{i,1} + L_{i,2} + \cdots + L_{i,k} \geq t$, where $k = \min_a\{a \geq 1 | L_{i,1} + L_{i,2} + \cdots + L_{i,a} \geq t\}$. The subscript i, t denotes that this is the auxiliary variable for the ith renewal process for the tth observation. Its value must be greater than or equal to t to ensure that the ith renewal process has run for at least t units of time. If $x_{i,t} = t$ then a renewal has taken place at t otherwise the next renewal is set to take place at some time after t, or more precisely, at time $x_{i,t}$. This auxiliary variable is required for each of the m independent copies of the renewal process, so we define $\boldsymbol{x_t} = (x_{1,t}, \ldots, x_{m,t})$. Furthermore, we have a set of N particles, so the algorithm must carry through a matrix of auxiliary parameters, $(\boldsymbol{x_t^1}, \ldots, \boldsymbol{x_t^N})$ where $\boldsymbol{x_t^j}$ is the vector of auxiliary parameters for the jth particle.

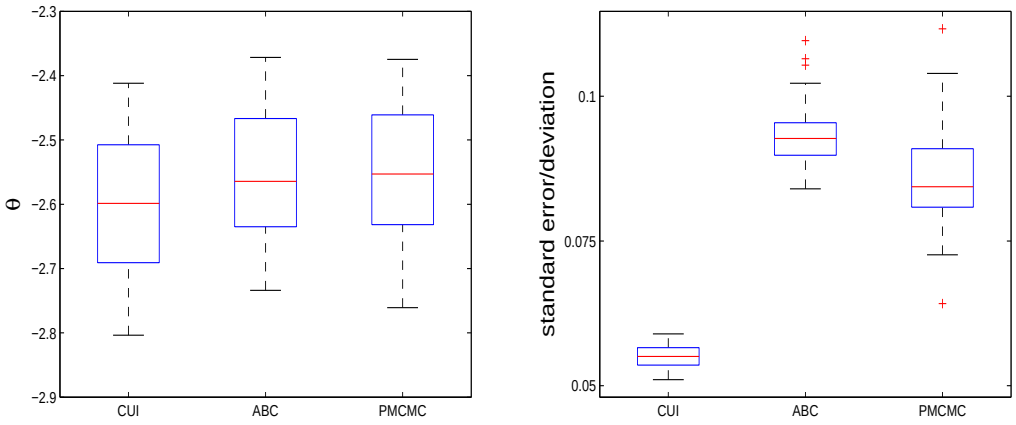
For this model it is possible that a particle is unable to simulate the next observation based on the value of its auxiliary variable. For example if $\sum_{i=1}^m 1(x_{i,t} = t + 1) > y_{t+1}$ then for a particular particle the number of renewals set to take place at time $t + 1$ is already greater than the next observation y_{t+1}, and thus cannot generate the data point. Alternatively if $\sum_{i=1}^m 1(x_{i,t} \leq t + 1) < y_{t+1}$ then there are too many renewals set to take place after $t + 1$ relative to the observed y_{t+1}. Therefore we have an indicator for the jth particle at time t, I_t^j, specifying whether or not the particle has a non-zero probability of being able to generate the next observation, y_{t+1}. This expression is given by

$$I_t^j = 1\left(\sum_{i=1}^m 1(x_{i,t} = t + 1) \leq y_{t+1} + \epsilon_{t+1}\right) \times 1\left(\sum_{i=1}^m 1(x_{i,t} \leq t + 1) \geq y_{t+1} - \epsilon_{t+1}\right).$$

To save on model simulations, in the particle filter we consider only particle indices $\mathcal{K}_t = \{j \in [1, N] | I_t^j = 1\}$ in the resample stage. The probability of accepting a simulation at time $t + 1$ is adjusted to be $\#\mathcal{K}_t/\text{sims}_{t+1}$ where sims_{t+1} is the total number of simulations required to produce N particles at time $t + 1$ in the particle filter. It is important to note that the above is not required to implement the method, it is merely for computational savings.

We found that for this model it was possible that $\#\mathcal{K}_t = 0$, implying the particle filter breaks down. This can be mitigated by increasing the number of particles. For the simulated data below, this was a rare occurrence. However, for the real data below with $\epsilon = 0$ it was rather common, however this may suggest a poor model choice (or a poor parameter proposal) rather than a blight on the methodology itself. When this occurred in the algorithm the proposal was rejected.

For the first example of this section we use simulated data based on a discretised Pareto lifetime distribution, $f_n \propto n^\theta$ for $n = 1, \ldots, 500$ with $L_0 = 0$. $T = 100$ observations were



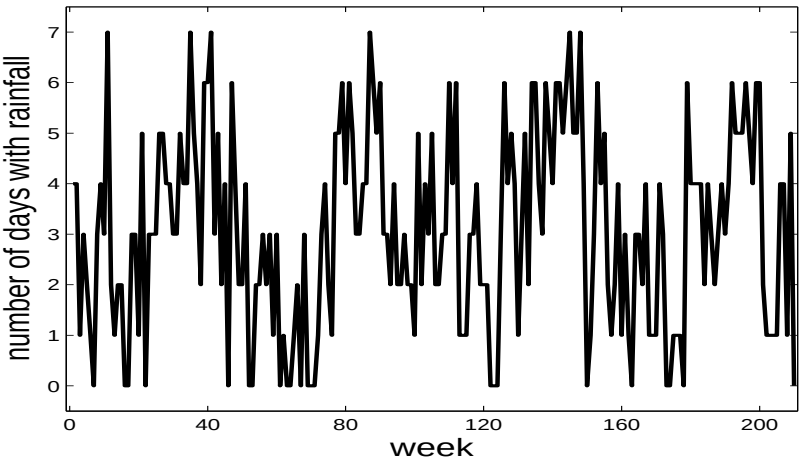
(a) parameter estimates

(b) uncertainty estimates

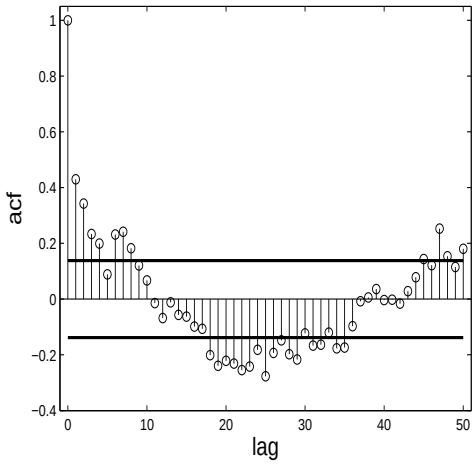
Figure 5: Boxplots of the (a) parameter estimates and (b) estimates of the uncertainty based on 50 independent datasets drawn from the binomial renewal process model with a discretised Pareto lifetime distribution (with $\theta = -2.5$) for three approaches: the approach of Cui and Lund (2009), ABC and PMCMC

simulated with a binomial marginal superpositioning $m = 10$ renewal processes (here $\theta = -2.5$). It should be noted that with this choice of parameter value, a dataset with long memory is created. 50 independent datasets were produced for analysis. We applied three methodologies: the frequentist estimation approach of Cui and Lund (2009), traditional ABC with summary statistics given by the sample mean and the first five autcorrelations and our PMCMC approach with exact matching. Boxplots of the posterior means (parameter estimates for Cui and Lund (2009)) and the posterior standard deviations (standard errors for Cui and Lund (2009)) of θ for the 50 different datasets is shown in Figure 5. It can be seen that the estimation approach of Cui and Lund (2009) produces a biased estimate. There are two issues with this estimation methodology: (1) such an estimation approach is only fully efficient for stationary Gaussian processes and (2) it assumes that the process has already reached stationarity. We are considering low count data with binomial marginals and thus a Gaussian approximation is not suitable. Furthermore assumption (2) is violated as the renewal process models are only asymptotically stationary, and will take a long time to reach stationarity if the process has long memory. ABC and PMCMC produce similar results but are still slightly biased (less so than the approach of Cui and Lund (2009)). It is evident from the boxplot in Figure 5(b) that PMCMC tends to provide a more precise estimate of θ. The results here for Cui and Lund (2009) suggest that the uncertainty in the parameter estimate is being somewhat underestimated. This is also not a surprising result, given the violation of the above mentioned assumptions (that the estimation approach of Cui and Lund (2009) is only fully efficient for stationary Gaussian time series).

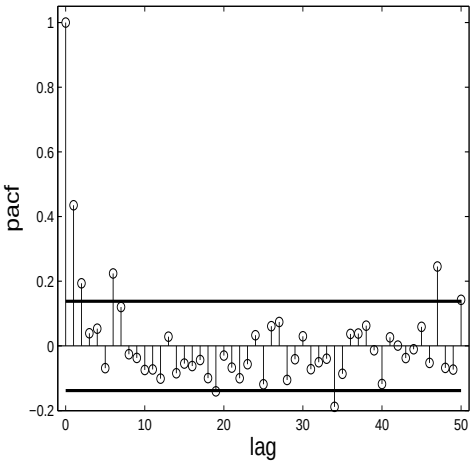
Finally, the algorithm is applied to some real data analysed in Cui and Lund (2009), which is the number of rainy days per week in Key West, Florida, over a period of $T = 210$ consecutive weeks. The data is shown in Figure 6. Cui and Lund (2009) propose a binomial marginal



(a) data



(b) acf



(c) pacf

Figure 6: (a) Number of rainy days per week in Key West, Florida for $T = 210$ consecutive weeks. (b) Autocorrelation function. (c) Partial autocorrelation function.

with a mixture of two geometric distributions (starting at 1) with different success probability parameters for the lifetime distribution. There are three parameters in the model: the two geometric parameters, p_1 and p_2, and the mixture weight parameter, w. For identifiability we restrict $p_1 \in (0, 0.5)$, $p_2 \in (0.5, 1)$ and $w \in (0, 0.5)$ (in the Bayesian analysis we use uniform priors over these intervals). The mixture of geometric distributions creates a process with short memory and requires little time to reach stationarity (indeed, a geometric lifetime distribution creates a stationary process due to the memoryless property of the geometric distribution).

Here we assume that the data has been observed from a process that has already reached steady state (which the estimation methodology of Cui and Lund (2009) implicitly assumes). Following the approach of Cui and Lund (2009), we obtained almost identical results to theirs. The parameter estimates with standard deviations were: $p_1 = 0.1245$ (0.0203), $p_2 = 0.7769$ (0.0286) and $w = 0.1492$ (0.0315).

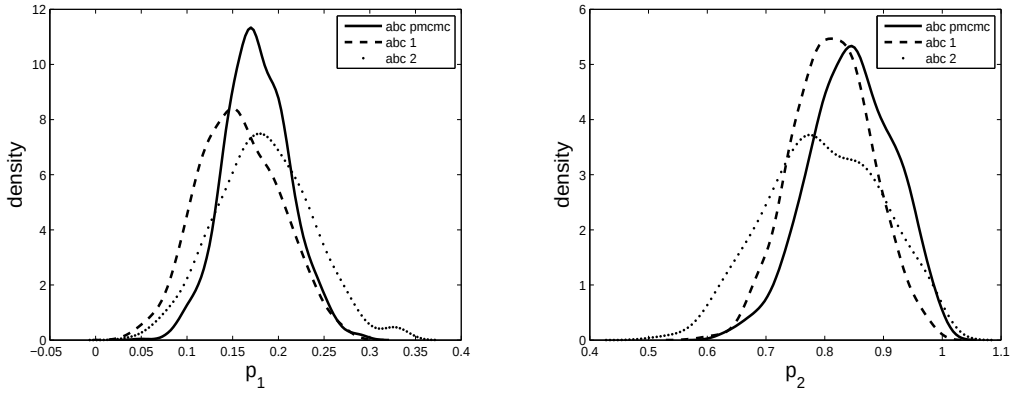
For each iteration in our approach we used a burn-in for the renewal process model of 100 observations (which was sufficient to achieve stationarity), and therefore obtain a distribution for the auxiliary parameters at time 0, $(\boldsymbol{x_0^1}, \ldots, \boldsymbol{x_0^N})$. Again $N = 1000$ particles were used. It quickly became apparent that the model is unable to explain the data, in contrast to the conclusion of Cui and Lund (2009), who suggest that the model is a reasonable fit to the data. Hence a tolerance value of $\epsilon_t = 1$ was used for all t except $t = 179$, in which case $\epsilon_{179} = 2$. The model cannot explain the immediate jump from $y_{178} = 0$ to $y_{179} = 6$. The (approximate) posterior distributions based on our approach can be found in Figure 7.

Due to the introduction of the tolerance, our method only obtains approximate inferences. Hence it seems sensible to compare it with other approximations, such as traditional ABC. The same burn-in was used when simulating data. We considered two sets of summary statistics: (1) mean, first five autocorrelations and first two partial autocorrelations, (2) mean, standard deviation, first five autocorrelations and first two partial autocorrelations. The two posteriors are shown in Figure 7. The ABC posteriors based on summary 1 provide point estimates in closer agreement with Cui and Lund (2009). When including the standard deviation into the set of summaries it became evident that the model underestimates this feature of the model. Therefore a high discrepancy was obtained with this set of summaries. It appears in this case the ABC inference results in less bias compared with our method of introducing the tolerance. Point estimates based on our approach did not recover the autocorrelations as well as the ABC approach. Again the Cui and Lund (2009) method produces standard errors which are substantially less than any of the posterior standard deviations.

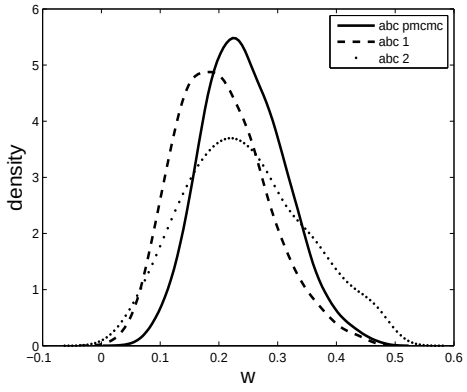
In the above examples each ABC run was quite fast compared with our algorithm. However, performing the ABC algorithm over different choices of the summary statistics was cumbersome. In appendix B we apply a different model that can explain the variability better.

4 Discussion

Here a general Bayesian methodology was presented to perform inference on parameters of models for low count time series data with intractable likelihoods. The method can work on partially observed processes and non-Markovian models through the use of auxiliary variables



(a) p_1 (b) p_2



(c) w

Figure 7: Posterior densities for when the mixture of geometric distributions lifetime model fitted to the rainfall data in Figure 6. Posteriors for the parameters are shown for our method as well as for ABC fits with summary statistics (see text for more details).

in the particle filter. Due to the sampler being on the space of $\boldsymbol{\theta}$ alone, efficient reversible jump proposals can be developed straightforwardly, which leads to robust fully-Bayesian model comparisons. When the observed data cannot be matched exactly, the algorithm allows for ABC inference by introducing a tolerance. No summary statistics are required. Our algorithm represents an alternative to Neal and Subba Rao (2007) and Enciso-Mora et al. (2009) for INARMA models as our algorithm is on the marginal space rather than the augmented space. Our method is the first likelihood-based estimation methodology proposed for the renewal process models in Cui and Lund (2009).

The method of Barthelmé and Chopin (2012) is another summary statistic free ABC method that may be applicable to the models considered in this paper. The method is very fast as it is based on an expectation propagation approximation. However, the method is restricted to posterior distributions that can be well described by a distribution in the exponential family (an assumption that cannot be investigated for complicated inference problems). Furthermore, the method of Barthelmé and Chopin (2012) requires further development if some of the variables are unobserved or if the process is non-Markovian. The unobserved variables could be included in the parameter vector $\boldsymbol{\theta}$, however the assumption that the posterior distribution for each of the parameters in $\boldsymbol{\theta}$ is well described by a distribution from the exponential family is unlikely to hold.

As was seen in two of the examples, the method can fail if there are outliers (with respect to a particular model). The algorithm can become stuck constantly trying to generate data from the model to match the observed outlier. However, if an outlier exists it may be that there is sufficient reason to remove such a data point or possibly that the model is not an adequate representation of the true underlying process. However, the method we propose can be used to identify if a model is not appropriate for the data and identify outliers in the data. For example, it was quickly realised that the mixture of geometric distributions renewal model was not able to generate the observed rainfall data (Figure 6) exactly (a point not realised by Cui and Lund (2009)). A real advantage of these methods is that they have a built in predictive check so that poorly fitting models are discovered quickly.

The method may be applicable to data with large counts or continuous data, but would require some development. Clearly, exact matching will be computationally prohibitive in such situations. Therefore, ABC will be required, potentially with a different discrepancy function (for example, a discretised Gaussian kernel) utilised here that is more appropriate for large values. The variance of the data at some time points may be greater than others, and this needs to be considered. We are currently exploring this in other research. Partially observed data and non-Markovian models could be handled in the same way via auxiliary variables in the particle filter. There are, of course, situations where models produce near chaotic data (see, for example, Wood (2010)), and methods which attempt to match simulated with observed datasets are not feasible.

Acknowledgements

The authors are grateful to Candice Hincksman for useful discussions on the model of Cui and Lund (2009).

References

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342.

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725.

Bailey, N. T. J. (1975). *The mathematical theory of infectious diseases and its applications*. Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE.

Barthelmé, S. and Chopin, N. (2012). Expectation-propagation for likelihood-free inference. *arXiv:1107.5959v2*.

Becker, N. G. (1989). *Analysis of infectious disease data*, volume 33. Chapman & Hall/CRC.

Blum, M. G. B., Nunes, M. A., Prangle, D., and Sisson, S. A. (2012). A comparative review of dimension reduction methods in approximate Bayesian computation. *To appear in Statistical Science*.

Cappellari, L. and Jenkins, S. P. (2003). Multivariate probit regression using simulated maximum likelihood. *The Stata Journal*, 3(3):278–294.

Cui, Y. and Lund, R. (2009). A new look at time series of counts. *Biometrika*, 96(4):781–792.

Drovandi, C. C. and Pettitt, A. N. (2008). Multivariate Markov process models for the transmission of Methicillin-resistant *Staphylococcus aureus* in a hospital ward. *Biometrics*, 64(3):851–859.

Eduarda Silva, M. and Pereira, I. (2012). Detection of additive outliers in Poisson integer-valued autoregressive time series. *arXiv:1204.6516v1*.

Enciso-Mora, V., Neal, P., and Subba Rao, T. (2009). Efficient order selection algorithms for integer-valued ARMA processes. *Journal of Time Series Analysis*, 30(1):1–18.

Flury, T. and Shephard, N. (2011). Bayesian inference based only on simulated likelihood: particle filter analysis of dynamic economic models. *Economic Theory*, 27:933–956.

Freeland, R. K. and McCabe, B. P. M. (2004). Forecasting discrete valued low count time series. *International Journal of Forecasting*, 20(3):427–434.

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *Radar and Signal Processing, IEE Proceedings F*, volume 140, pages 107–113.

Jasra, A., Lee, A., Yau, C., and Zhang, X. (2013). The alive particle filter. *http://arxiv.org/abs/1304.0151*.

Le Gland, F. and Oudjane, N. (2006). *Lecture Notes in Control and Information Sciences*, chapter A Sequential Particle Algorithm that Keeps the Particle System Alive. Stochastic Hybrid Systems: Theory and Safety Critical Applications, pages 351–389. Number 337. Springer.

Martin, V. L., Tremayne, A. R., and Jung, R. C. (2011). Efficient method of moments estimators for integer time series models.

Moler, C. and van Loan, C. (2003). Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1):3–49.

Neal, P. and Subba Rao, T. (2007). MCMC for integer-valued ARMA processes. *Journal of Time Series Analysis*, 28(1):92–110.

O'Neill, P. D. and Roberts, G. O. (1999). Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(1):121–129.

Pericchi, L. R. (2005). *Bayesian thinking: modeling and computation*, chapter Model selection and hypothesis testing based on objective probabilities and Bayes factors, pages 115–149. Elsevier B. V.

Sidje, R. B. (1998). Expokit: a software package for computing matrix exponentials. *ACM Transactions on Mathematical Software (TOMS)*, 24(1):130–156.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. H. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface*, 6(31):187–202.

White, S. R., Kypraios, T., and Preston, S. P. (2013). Fast approximate Bayesian computation for discretely observed Markov models using a factorised posterior distribution. *arXiv:1301.2975v1*.

Wilkinson, R. D. (2008). Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *arXiv:0811.3355*.

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104.

Zhang, H., Wang, D., and Zhu, F. (2011). The empirical likelihood for first-order random coefficient integer-valued autoregressive processes. *Communications in Statistics - Theory and Methods*, 40(3):492–509.

Appendix A

Here we demonstrate another application of model selection for INARMA models. This example is to the number of monthly claimants of burn related injuries to the Workers Compensation Board, British Columbia, Canada. The data spans between January 1984 and December

1994, and is analysed in Freeland and McCabe (2004). The data is shown in Figure 8(a). From the aucorrelation function (Figure 8(b)) and partial autocorrelation function (Figure 8(c)) of this data one may entertain an INAR(1) or INAR(2) model. We perform within model inference as well as use reversible jump to estimate the posterior model probabilities.

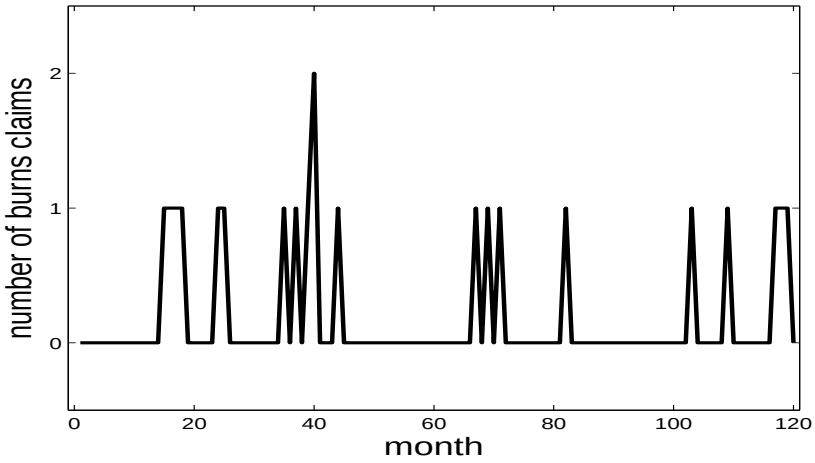
The INAR(1) and INAR(2) models were fitted separately using 21000 iterations of Algorithm 1 with a normal random walk. The random walk standard deviations of the AR parameters were 0.1 and 0.03 for λ. The first 1000 iterations were discarded as burn-in and the resulting sample was thinned by a factor of 10. We match the data exactly, so obtaining exact Bayesian inferences and use $N = 200$ particles. Note that the exact likelihood can be obtained for the INAR(1) straightforwardly but for illustrative purposes we use the simulated likelihood. In Figure 9 the inferences on the INAR(1) model are compared with the simulated and exact likelihoods for validation. For the exact likelihood results the same MH algorithm (with the true likelihood replacing the simulated likelihood) was used with a thinning factor of 10 for the density estimation.

Gamma distributions were fitted via maximum likelihood to the posterior samples. These gamma distributions fit the posteriors reasonable well but importantly cover the tails. These are used as independent proposals of a reversible jump algorithm. In each iteration of the RJMCMC an independent proposal is made to the opposite model. If the move is accepted then the algorithm jumps to the other model, otherwise it remains in the same model. We use the RJMCMC output to obtain both within-model parameter inferences and estimated posterior model probabilities via the proportion of time the sampler spends in a particular model. We start in the INAR(1) model and use a single posterior sample from the associated within-model run so no burn-in is required. The algorithm was run for 20000 iterations and parameter densities are estimated from using a thinning factor of 5. The resulting posterior distributions can be seen in Figure 9. The posterior probability of the INAR(2) model was estimated to be 0.56, suggesting weak evidence in favour of this model.

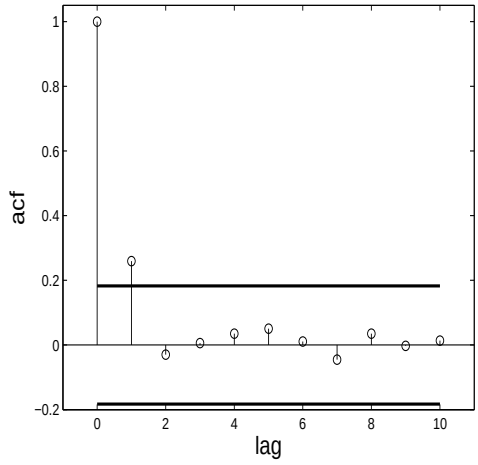
Appendix B

In an attempt to obtain a model with a better fit to the rainfall data (Figure 6), the marginal distribution was changed to a truncated negative binomial distribution with parameters $T = 10$ (the truncation number), $r = 2$ (the number of 'successes' required) and success probability p. Here the marginal can be interpreted as the number of failures until two successes, truncated at 10 trials. This marginal provides extra variability, at the expense of possibly generating an observation greater than 7. In this model there are 4 parameters for inference, p_1, p_2, w and p. In the particle filter we ran the $T = 10$ renewal processes sequentially and drew a (truncated) negative binomial random variate, independently for each proposal, to determine how many renewal processes to sum. We use the same auxiliary information as above. However, we do not use the indicators as in the main paper since more care is required as the number of renewal processes being superpositioned is random. Again $N = 1000$ particles was used but this time we were able to set $\epsilon = 0$. We also ran a traditional ABC analysis with summary statistics given by the mean, standard deviation, first five autocorrelations and first two partial autocorrelations.

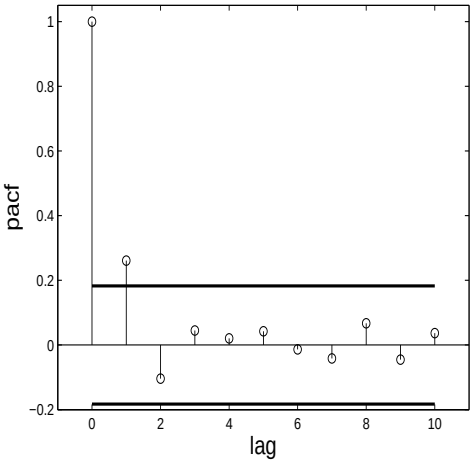
The posterior distributions are shown in Figure 10 for the two approaches. The ABC analysis



(a) data

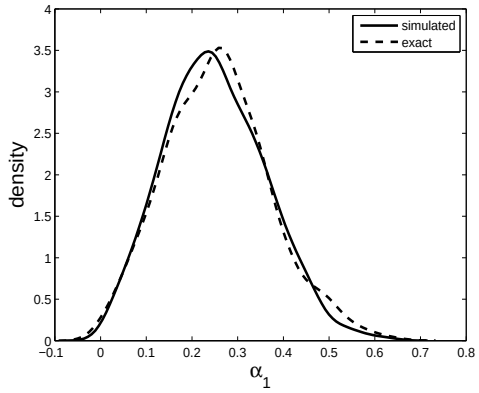


(b) acf

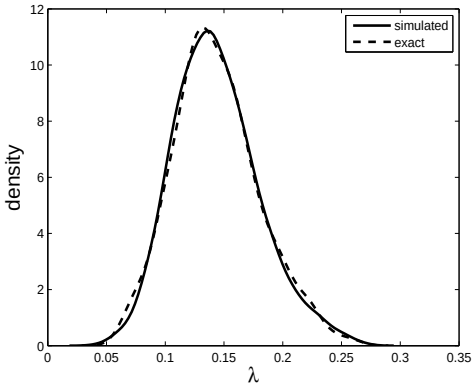


(c) pacf

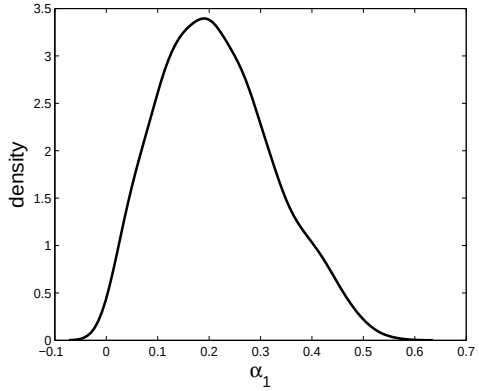
Figure 8: (a) The number of monthly burns claimants to the Workers Compensation Board, British Columbia, Canada, between January 1984 and December 1994. (inclusive). (b) Autocorrelation function. (c) Partial autocorrelation function.



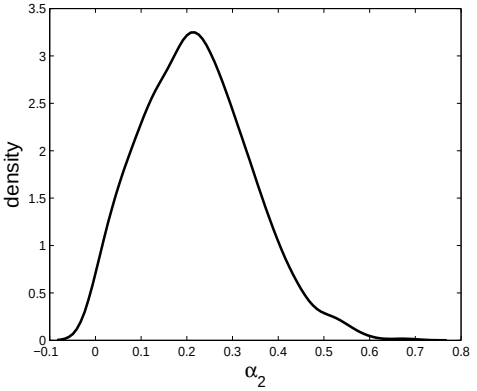
(a) α_1, INAR(1)



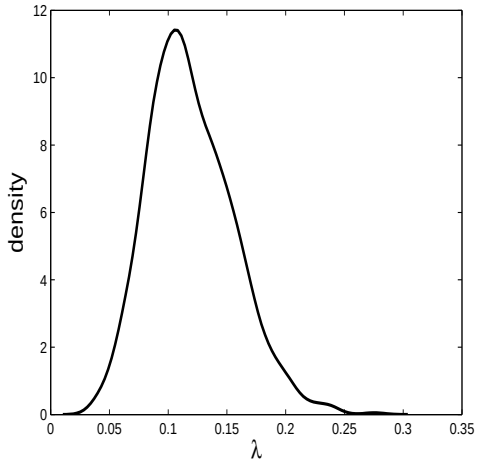
(b) λ, INAR(1)



(c) α_1, INAR(2)

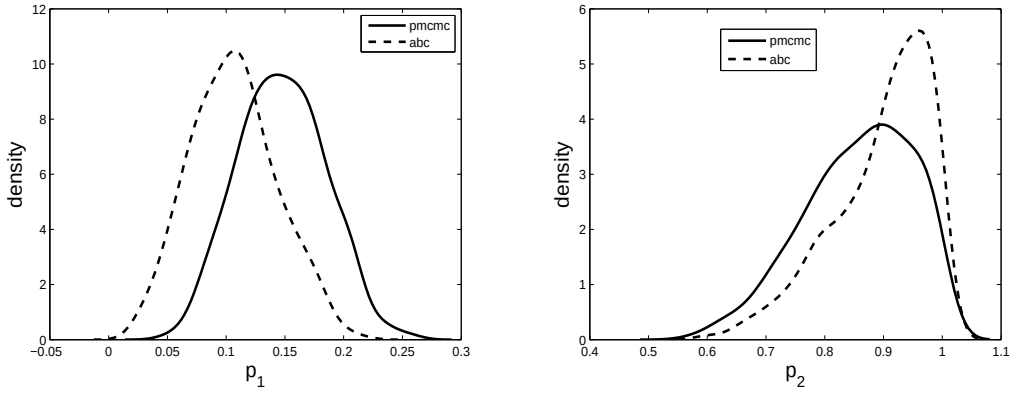


(d) α_2, INAR(2)

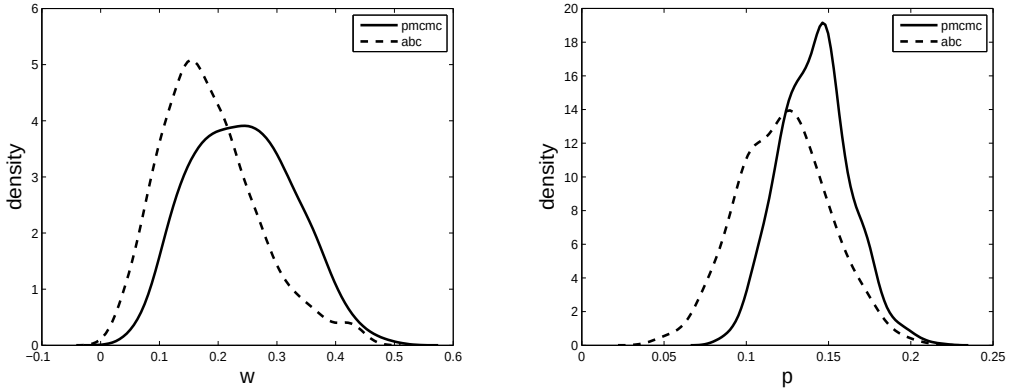


(e) λ, INAR(2)

Figure 9: Posterior densities for when the INAR(1) and INAR(2) models are fitted to the burns data in Figure 8(a). (a) and (b) Posteriors for INAR(1) model. (c), (d) and (e) Posteriors for INAR(2) model.



(a) p_1 (b) p_2



(c) w (d) p

Figure 10: Posterior densities for when the mixture of geometrics lifetime model fitted to the rainfall data in Figure 6 with a negative binomial marginal. Posteriors for the parameters are shown for our method as well as to an ABC fit with summary statistics (see text for more details).

indicated that this model can explain the variability in the data much better than the binomial marginal (however there is still a slight underestimation). However, the binomial model recovers the first two autocorrelations and first partial autocorrelation more accurately.